W0016

RETRIEVAL METHOD FOR INFORMATION AND INFORMATION STORAGE DEVICE

Patent Number: JP4158478
Publication date: 1992-06-01
Inventor(s): TANAKA SHINICHI; others: 01
Applicant(s):: MATSUSHITA ELECTRIC IND CO LTD
Requested Patent: JP4158478
Application Number: JP19900285023 19901022
Priority Number(s):
IPC Classification: G06F15/40 ; G06K9/72
EC Classification:
Equivalents: JP2932667B2

Abstract

PURPOSE:To reduce retrieval leakage by performing retrieval by enlarging a retrieval character string essentially so as to compensate the incompletion of character recognition when a character string is retrieved from a character code obtained by performing character recognition on a document. CONSTITUTION:An information processing means 2 enlarges the range of the character string to be retrieved, and retrieves the character string that coincides with either the character strings whose range are enlarged as reading out character code information within a range instructed by s storage means 3. When the character string of document information inputted as image information is retrieved from the character code information to which the character recognition is applied in such way, the range can be enlarged essentially and the retrieval is performed by substituting the sum set of the character string and all the character strings with possibility to recognize the character string erroneously for the character string by referring to a correspondence table that is a table representing the tendency of erroneous recognition proper to the algorithm of character recognition. In such a way, it is possible to remarkably reduce the frequency of the retrieval leakage even when the erroneous recognition is performed by a character recognition means 5.

TOP

W0016

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平4−158478

⑤1Int.Cl.⁵	識別記号	庁内整理番号	㊸公開　平成4年(1992)6月1日
G 06 F 15/40	510 L	7056−5L	
G 06 K 9/72		7737−5L	

審査請求　未請求　請求項の数 3　(全6頁)

㊹発明の名称　情報の検索方法および情報蓄積装置

㉑特　願　平2−285023

㉒出　願　平2(1990)10月22日

㉗発明者　田中　伸一　大阪府門真市大字門真1006番地　松下電器産業株式会社内

㉗発明者　松川　茂　大阪府門真市大字門真1006番地　松下電器産業株式会社内

㉑出願人　松下電器産業株式会社　大阪府門真市大字門真1006番地

㉔代理人　弁理士　小鍜治　明　外2名

明細書

1. 発明の名称

情報の検索方法および情報蓄積装置

2. 特許請求の範囲

(1)画像情報として入力された情報を文字認識して得られた結果を蓄積した文字コード情報から所定の条件式に合致した文字列を検索するときに、上記認識のアルゴリズムに固有の不完全性を補うように検索の範囲を拡大して検索することを特徴とする情報蓄積装置のための情報の検索方法。

(2)画像情報として入力される文書情報を蓄積する画像蓄積手段と、上記文書情報に含まれる文字を認識する認識手段と、この認識手段から出力されるコード情報を蓄積する補助情報蓄積手段と、検索範囲として入力される条件式に基づいて上記補助情報蓄積手段に蓄積されたコード情報内で検索する検索手段とを具備し、

上記検索手段は、上記認識手段の不完全性を補うように検索範囲を実質的に拡大して検索することを特徴とする情報蓄積装置。

(3)検索手段は、認識手段の認識のアルゴリズムに固有の誤認識の傾向に基づき、各文字に対してこれを誤認識し易い他の文字を対応づける対応表を有し、検索条件式に含まれる文字列の各文字を、この文字と、上記対応表で関連付けられた誤認識し易い他の文字との和集合に置き換えて検索することを特徴とする特許請求の範囲第2項記載の情報蓄積装置。

3. 発明の詳細な説明

産業上の利用分野

本発明は、画像情報として入力される文書を電子的に蓄積する情報蓄積装置と情報の検索方法に関するものである。

従来の技術

近年、文書や図面を画像情報として入力してこれを電子的に蓄積する文書ファイル装置と呼ばれる情報蓄積装置が、文書や図面の管理を専業とする部署を中心に普及し始めている。

以下、図面を参照しながら、上述した従来の情報蓄積装置の一例について説明する。

第5図は従来の情報蓄積装置の動作を示すフローチャートである。第5図a)は文書の登録時、第5図b)は検索時の動作をそれぞれ示すものである。

以上のように動作する情報蓄積装置について、以下その動作をさらに詳しく説明をする。

まず、文書を登録するときには、イメージスキャナなどの画像入力装置で、文書情報を画像情報として取り込む。取り込んだ画像情報は光ディスク装置などの記憶装置に蓄積される。続いて、蓄積した画像情報の文書名、分類、作成者、キーワードなどの検索に利用する補助情報をキーボードから入力し、この補助情報に対応する画像情報を示す情報を付加して所定の場所に記憶する。

このようにして蓄積された画像情報を検索するときには、キーボードから補助情報を限定する検索条件を入力し、所定の場所に記憶された補助情報の中でこれに合致する補助情報を検索する。このようにして検索しようとする文書情報の補助情報が特定されると、これに対応する文書を読み出すことができる。（例えば、オーム社「オフィスオートメーション入門」111～113ページ）

発明が解決しようとする課題

しかしながら上記のような動作では、文書を登録するときに必ず検索のための補助情報を入力する必要があり、登録に手間がかかるばかりでなく、複数の人で文書を登録したり検索したりするときには、各人の間でキーワードの整合性や一貫性をとる必要もあり、キーワード体系の管理が大変であるという問題点を有していた。

本発明は上記問題点に鑑み、文書を登録するときに、検索のための補助情報をわざわざ入力しなくても後で検索することが可能な情報の検索方法および情報蓄積装置を提供するものである。

課題を解決するための手段

上記課題を解決するために、本発明の情報の検索方法および情報蓄積装置は、画像情報として入力される文書情報から文字を認識した文字コード情報から所定の文字列を検索しようとするもので、検索すべき文字列を含む文書の画像情報からこの文字列を認識するとき、認識のアルゴリズムに付随する不完全性のためにこの文字列を誤認識する可能性のある他の文字列と正しい文字列のいずれかに合致する文字列を検索するようにしたものである。

作　用

本発明は上記した方法によって、文書の画像情報を文字認識して得られる文字コード情報の中で所定の文字列を認識するので、文書情報に検索用のキーワードなどの補助情報を付加しなくても直接文書情報から検索することが可能であり、文字認識の不完全性を補うように検索の条件を拡大するので、誤認識に伴う検索漏れを回避することが可能となる。

認識の不完全性を補う方法について、その原理をさらに説明する。

第3図は、理想的に文字認識できる場合を示す概念図である。同図において、実線で囲んだ領域a～領域hは、それぞれ、仮想的な文字a～文字hのパターンの存在範囲を示すもので、破線で囲んだ領域A～領域Bは、それぞれ、文字a～文字hと認識されるパターンの範囲を示すものである。この場合には、領域a～領域hは、それぞれ、領域A～領域Hに完全に包含されており、文字a～文字hがすべて正しく認識されることは明らかである。

一方、第4図は認識が正しく行われない場合を示す概念図である。第3図の場合と同様に、実線で囲んだ領域i～領域pは、それぞれ、仮想的な文字i～文字pのパターンの存在範囲を示し、破線で囲まれた領域I～領域Pは、それぞれ、文字i～文字pと認識されるパターンの範囲を示すものである。なお、領域X～領域Zは、どの文字にも認識できない領域を示すものである。この場合には、すべての文字i～文字pが領域I～領域Pに完全に包含されているわけではなく、完全な文字認識を行うのは不可能である。例えば、文字iは、ほとんどの場合、文字iと正しく認識されるが、文字jや文字nに近いパターンで書かれていると、それぞれ文字jや文字nに誤認識されることとなる。また、文字mと文字oは、存在し得る

パターンの領域が重なっており、文脈などから意味を理解するような、パターン認識以外の手段を併用する以外に誤認識を避ける方法はない。このようなことは、異なる文字体系が混在する場合におこり得るものである。例えば、漢字の'入'とギリシャ文字の'λ'や、数字の'0'とアルファベットの'O'などがその好例である。

このような誤認識が、どの文字に対してどのように発生するかということは、認識アルゴリズムに固有の傾向を有しており、その傾向さえ把握できておれば、検索のときにその欠点を補うことが可能である。例えば、文字iが文字jに誤認識されたとしよう。この場合、その認識結果を印刷や表示などの形で出力すると支障があるが、文字iで検索する場合には、（文字i＋文字j＋文字n）で検索すれば検索漏れは回避される。検索範囲を拡大することによって、余分なものも検索されてしまうが、検索条件を変えて絞り込みを行えばほとんど支障はなくなる。また、実際には、1文字で検索することはほとんどなく、数文字を組み合わせた熟語で検索されるので、検索範囲は実質的にはそれほど極端に拡大されることはない。例えば、'入力'という文字列で検索する場合、'入'という文字を'入＋λ'に、'力'を'力＋λ'にそれぞれ拡大しても、文字列としては'入力＋λ力＋入λ＋λλ'に拡大されるだけで、'λ力'、'入λ'、'λλ'などはほとんど存在しないので、実質的な検索範囲の拡大は極わずかとなる。

このように、検索時に文字認識の不完全性を補うことによって、検索漏れという不都合な事態を大幅に減少させることが可能となる。

実　施　例

以下本発明の一実施例の情報の検索方法について、図面を参照しながら説明する。

第1図は本発明の第1の実施例における情報蓄積装置のブロック図を示すものである。第1図において、1は画像入力手段で、手書きあるいは印刷された文書から画像として情報を取り込む。2は情報処理手段で情報の入出力の制御やさまざまな処理を行う。3は記憶手段で、情報処理手段2の取り扱う情報を必要に応じて記憶する。4はコード入力手段で、画像入力手段1から入力された画像情報の補助情報や、検索のための文字列などを入力する。5は文字認識手段で、情報処理手段2から送られてくる画像情報から文字を切り出して認識し、文字認識して得られた文字コードを情報処理手段2に返す。6は出力手段で、コード入力手段4から入力される指示に従って、指示された特定の情報や指示に従って検索して抽出された情報などを出力する。

以上のように構成された情報蓄積装置について、以下、第1図および第2図を用いてその動作を説明する。

まず、第2図は本実施例における情報蓄積装置の動作のフローチャートを示したものであって、第2図a)は、文書情報を蓄積する場合、第2図b)は、所望する文書情報を検索する場合をそれぞれ示すものである。文書情報を蓄積するときには、まず、手書きあるいは印刷された文書を、イメージスキャナのような画像入力手段1で画像情報として読み取り、情報処理手段2に転送する。情報処理手段2は、この画像情報のフォーマットを整え、記憶手段3に転送してファイルとして記憶させる。さらに、必要に応じて、キーワードなどの補助情報をコード入力手段4から入力して、情報処理手段2はこの補助情報を所定のフォーマットに整えて、記憶手段3に送出して所定の場所に記憶させる。一方、情報処理手段2は、文字認識手段5にも情報画像情報を送出する。文字認識手段5は、この画像情報から文字を順次切り出して、これを認識し、文字コードに変換する。文字認識手段5は、認識して得た文字コード情報を情報処理手段2に返す。情報処理手段2は、この文字コード情報を所定のフォーマットに整えた後、記憶手段3に送出して、所定の場所に記録させる。

以上のようにして蓄積された文書情報から、所望の情報を検索するときには、まず、検索を行う対象とするファイルを限定するために必要に応じてキーワードなどによる制限条件をコード入力手段4から入力する。もちろん、すべてのファイル

を対象にするときには、このような制限条件の入力は必要としない。次に、検索すべき文字列を再びコード入力手段4から入力する。この文字列は情報処理手段2に転送され、情報処理手段2は内蔵する対応表に基づいて、この文字列の範囲を拡大する。この、文字列の範囲を拡大する過程を、さらに詳しく説明する。

対応表とは、文字認識手段5が各文字を認識するときに誤認識する可能性のある文字を、各文字に対応させた表である。例えば、第3図に示すように、文字ｉ～文字ｐのパターンの範囲および文字ｉ～文字ｐと認識されるパターンの範囲である領域Ｉ～領域Ｐが分布しているとすれば、これらの文字に関する対応表は表1のようになる。

（以下余白）

表1

文字ｉ	文字ｉ，文字ｊ，文字ｎ
文字ｊ	文字ｊ，文字ｎ，文字?
文字ｋ	文字ｋ
文字ｌ	文字ｌ
文字ｍ	文字ｍ
文字ｎ	文字ｎ
文字ｏ	文字ｏ，文字ｍ
文字ｐ	文字ｐ

この対応表は、実際にはそれぞれの文字を表す文字コードで構成されており、文字? は、認識できなかった文字に割り当てる特殊コードを意味する。

検索する文字列が‘文字ｊ文字ｍ文字ｏ’の3文字から成る文字列であるとすれば、文字ｊは文字ｊと文字ｎと文字? との和集合に置き換え、文字ｏは文字ｏと文字ｍとの和集合に置き換える。したがって、検索は

‘文字ｊ文字ｍ文字ｏ’
‘文字ｊ文字ｍ文字ｍ’
‘文字ｎ文字ｍ文字ｏ’
‘文字ｎ文字ｍ文字ｍ’
‘文字? 文字ｍ文字ｏ’
‘文字? 文字ｍ文字ｍ’

の6種類の文字列の和集合で行われることとなる。

さて、情報処理手段2は、以上のようにして、検索すべき文字列の範囲を拡大し、記憶手段3から、指定された範囲の文字コード情報を読み出しながら、範囲の拡大された文字列のいずれかと一致する文字列を検索する。情報処理手段2は検索によって抽出された文書情報を、出力手段6に送出する。出力手段6が、例えばＣＲＴの場合にはそれに表示され、プリンタの場合には、その情報が印刷される。

以上のように本実施例によれば、画像情報として入力された文書情報を文字認識した文字コード情報から文字列を検索するとき、文字認識のアルゴリズムに固有の誤認識の傾向を表すテーブルである対応表を参照して、上記文字列を、この文字列とこれを誤認識する可能性のあるすべての文字列との和集合に置き換えることによって、実質的に範囲を拡大して検索することにより、例え文字認識手段5が誤認識しても、検索漏れの頻度を大幅に減少させることができる。

なお、上記の実施例においては、対応表に基づいて検索文字列を、誤認識し易い他の文字列との和集合に置き換えて検索するように構成したが、本発明の主旨は、文字認識した文字コードを検索にだけ用いるときには、検索のときに、認識の不完全性を補うようにすれば、誤認識は大きな問題とはならないことに着眼して、検索すべき文字列を、認識の不完全性を補うように実質的に文字列の範囲を拡大して検索しようとするものである。したがって、検索する文字列の範囲を実質的に拡大する手段はどのような手段であっても特に限定されるものではない。

発明の効果

以上のように本発明は、文書を文字認識して得

られた文字コードから文字列を検索するとき、文字認識の不完全性を補うように検索の文字列を実質的に拡大して検索することによって、検索漏れの頻度を大幅に減少させることができる。

4. 図面の簡単な説明

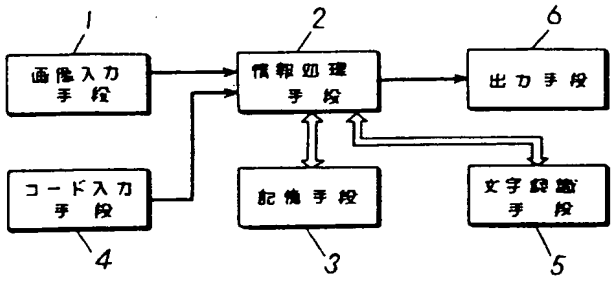
第1図は本発明の一実施例における情報蓄積装置のブロック図、第2図は上記実施例における情報蓄積装置の動作を示すフローチャート、第3図は理想的な文字認識の場合を示す概念図、第4図は不完全な文字認識の場合を示す概念図、第5図は従来の情報蓄積装置の動作を示すフローチャートである。

2……情報処理手段、3……記憶手段、4……コード入力手段、5……文字認識手段。

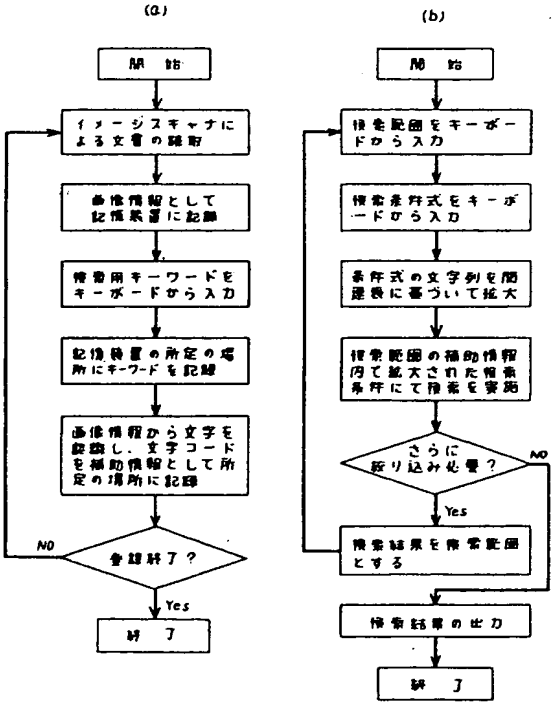
代理人の氏名　弁理士　小鍜冶　明

ほか2名

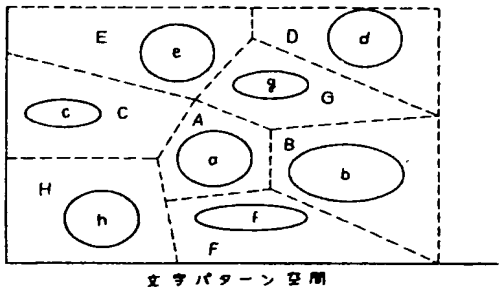
第1図



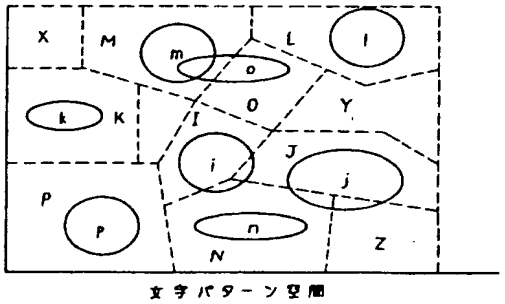
第2図



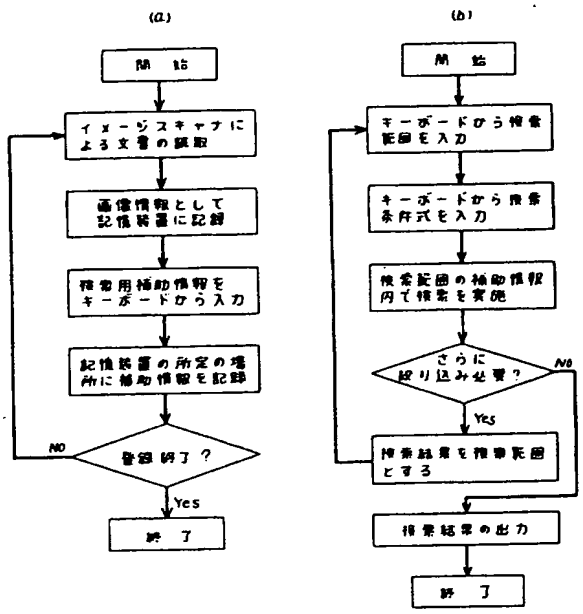
第3図



第4図



第 5 図


(a)
開 始
イメージスキャナによる文書の読取
画像情報として記憶装置に記録
検索用補助情報をキーボードから入力
記憶装置の所定の場所に補助情報を記録
登録終了？
NO
Yes
終 了
(b)
開 始
キーボードから検索範囲を入力
キーボードから検索条件式を入力
検索範囲の補助情報内で検索を実施
さらに絞り込み必要？
NO
Yes
検索結果を検索範囲とする
検索結果の出力
終 了

SPECIFICATION

1. Title of the Invention

Information Retrieval Method and Information Storage Apparatus

2. Scope of Claims

(1) An information retrieval method for an information storage apparatus wherein when a character string matching a predetermined condition equation is retrieved from stored character code information obtained through recognition of characters input as image information, a retrieval range is broadened in order to compensate for incompleteness inherent in a character recognition algorithm.

(2) An information storage apparatus comprising: image storage means for storing document information input as image information; recognition means for recognizing characters contained in the document information; auxiliary information storage means for storing code information output from said recognition means; and retrieval means for retrieving the code information in accordance with a condition equation input as a retrieval range, wherein said retrieval means retrieves the code information by broadening a substantial retrieval range so that incompleteness of said recognition means can be compensated.

(3) An information storage apparatus according to claim 2, wherein said retrieval means has a table storing a correspondence between each character and another character or characters easy to be erroneously recognized

from the first mentioned character, the table being formed in accordance with erroneous recognition tendency inherent in a recognition algorithm to be used by said recognition means, and said retrieval means retrieves a character string contained in the retrieval condition equation as a sum-set of each character of the character string and the other character or characters which are easy to be erroneously recognized from the first mentioned character and are related by the correspondence table.

3. Detailed Description of the Invention

Industrial Application Field

The present invention relates to an information storage apparatus for electronically storing a document input as image information and to an information retrieval method.

Related Art

An information storage apparatus called a document filing apparatus for electronically storing documents and drawings input as image information in prevailing in business departments which mainly manage documents and drawings.

With reference to the accompanying drawings, an example of a conventional information storage apparatus will be described.

Fig. 5 is a flow chart illustrating the operation of a conventional information storage apparatus. Fig. 5(a) illustrates a document registration operation, and Fig. 5(b) illustrates a document retrieval operation.

The operation of the information storage apparatus operating as illustrated in Figs. 5(a) and 5(b) will be

detailed hereinunder.

When a document is registered, document information is captured as image information by using an image input apparatus such as an image scanner. The captured image information is stored in a storage device such as an optical disc. Next, auxiliary information to be used for retrieval, such as the document name, classification, author and keywords respectively of the stored image information, is input from a keyboard. The auxiliary information and information representative of the image information are stored in the storage device at a predetermined location.

When the image information stored in the storage device is retrieved, a retrieval condition for identifying the auxiliary information is input from the keyboard to retrieve the auxiliary information stored in the storage device and matching the retrieval condition. After the auxiliary information of the document information to be retrieved is retrieved, the corresponding document can be read (for example, refer to "Office Automation Guide" Ohm Co. pp. 111-113).

Problems to be Solved by the Invention

With the above operation, however, it is essential to enter the auxiliary information for retrieval when a document is registered. It takes labor to register a document. Furthermore, if a plurality of persons register and retrieve documents, it is necessary to have integrity and consistency of keywords used by those persons. Management of a keyword system becomes complicated.

Under the circumstances of such problems, the invention provides an information retrieval method and an information storage apparatus capable of retrieving a document without entering auxiliary information for retrieval when the document is registered.

Means for Solving the Problems

In order to solve the above problems, according to the retrieval method and information storage apparatus of the invention, a desired character string is retrieved from character code information obtained through recognition of characters of a document input as image information, wherein when the character string to be retrieved is recognized from the image information of the document, a character string matching either a target correct character string or another character string or strings having a possibility of being erroneously recognized from the target correct character string because of incompleteness of a recognition algorithm.

Operation

According to the information retrieval method of the invention, a desired character string is recognized from character code information obtained through character recognition of a document input as image information. Accordingly, a character string can be retrieved directly from the document information without adding retrieval auxiliary information such as keywords to the document information. Since a retrieval condition is broadened in order to compensate for incompleteness of character

recognition, it is possible to avoid a retrieval omission by erroneous recognition.

The principle of a method of compensating for incompleteness of recognition will be described.

Fig. 3 is a conceptual diagram illustrating ideal character recognition. In Fig. 3, areas a to h surrounded by solid lines indicate existence areas of patterns of virtual characters a to h, respectively. Areas A to H surrounded by broken lines indicate areas of patterns recognized as the characters a to h, respectively. Since the areas a to h are completely included in the areas A to H, respectively, it is obvious that the characters a to h are all recognized correctly.

Fig. 4 is a conceptual diagram illustrating the case that characters are not recognized correctly. Similar to Fig. 3, areas i to p surrounded by solid lines indicate existence areas of patterns of virtual characters i to p, respectively. Areas I to P surrounded by broken lines indicate areas of patterns recognized as the characters i to p, respectively. No character is recognized from areas X to Z. In this example shown in Fig. 4, all characters i to p are not completely included in the areas I to P so that perfect character recognition is impossible. For example, although the character i is recognized as the character i in some cases, if the character i is written by patterns similar to character j or n, the character i is erroneously recognized as the character j or n. The patterns of characters m and O are overlapped so that

erroneous recognition is inevitable unless means other than the pattern recognition, such as recognition of meaning from the context of a passage, is used together with the pattern recognition. Such a case occurs if different character systems are used. Typical examples are a kanji character "入" and a Greek character "ι", and a numeral "0" and an alphabet "O".

There is a tendency inherent in a recognition algorithm that such erroneous recognition occurs at which character in what manner. If this tendency can be grasped, such deficiency can be compensated during retrieval. Consider for example that the character i is erroneously recognized as the character j. In this case, although it is not suitable for printing or displaying, a search omission can be avoided if characters (i + j + n) are searched at the same time. If the search range is broadened, unnecessary characters are retrieved. However, if the retrieval condition is narrowed down there is no practical problem. In practice, retrieval using one character hardly occurs and retrieval using a compound word made of a combination of several characters is usually performed. Therefore, the retrieval range is not substantially broadened too much. Consider for example that a character string "入力" is retrieved. In this case, if the character "入" is broadened to "入 + ㇒ " and the character "力" is broadened to "力 + ｶ", the character string is broadened to "入力 + 入力+ ㇒ｶ". However, in this case, "㇒力", "入ｶ", "㇒ｶ" and the like hardly exist so that a substantial extension of the retrieval range

is very small.

As above, by compensating for incompleteness of character recognition during retrieval, search omissions can be reduced considerably.

Embodiment

An information retrieval method according to an embodiment of the invention will be described with reference to the accompanying drawings.

Fig. 1 is a block diagram showing an information storage apparatus according to a first embodiment of the invention. In Fig. 1, an image input unit 1 captures images of a hand-written or printed document. An information processing unit 2 performs information input/output control and various processes. A storage device 3 stores, when necessary, information to be used by the information processing unit 2. A code input unit 4 inputs auxiliary information of the image information input from the image input unit 1, a character string to be retrieved, and the like. A character recognition unit 5 cuts characters off the image information supplied from the image processing unit 2, recognizes them, and returns the obtained character codes back to the information processing unit 2. An output unit 6 outputs specific information designated by the inputs of the input unit 4, information retrieved and extracted in accordance with the inputs of the input unit 4, and other information. The operation of the information storage apparatus constructed as above will be described with reference to Figs. 1 and 2.

Fig. 2 is a flow chart illustrating an operation of the information storage apparatus of the embodiment. Fig. 2(a) illustrates a document registration operation, and Fig. 2(b) illustrates a document retrieval operation. When a hand-written or printed document is registered, the document is read with the image input unit 1 such as an image scanner as image information. The image information is transferred to the information processing unit 2. The information processing unit 2 changes the format of the image information and transfers the image information to the storage device 3 to be stored as a file. When necessary, auxiliary information such as a keyword is entered from the code input unit 4. The information processing unit 2 changes the format of the auxiliary information and transfers the auxiliary information to the storage device 3 to be stored at a predetermined location. The information processing unit 2 also transfers the image information to the character recognition unit 5 which sequentially cuts characters from the image information, recognizes them and converts them into character codes. The character recognition unit 5 returns the recognized and obtained character codes to the information processing unit 2. The information processing unit 2 changes the character code information to have a predetermined format, and sends the character code information to the storage device 3 to be stored at a predetermined location.

In retrieving desired information from the document information registered in the above manner, first, a

limitation condition is input from the code input unit 4 in order to limit a retrieval target file. The limitation condition may be written by a keyword. This limitation condition is not entered if all files are used as retrieval target files. Next, a character string to be retrieved is input from the code input unit 4. This character string is transferred to the information processing unit 2. In accordance with a correspondence table in the information processing unit 2, the information processing unit 2 broadens the range of the character string. A process of broadening the range of a character string will be described more in detail.

The correspondence table is a table storing a correspondence between each character and a corresponding character or characters having a possibility of being erroneously recognized as the first mentioned character by the character recognition unit 5. Assuming that the areas of patterns of the characters i to p and the areas I to P of patterns recognized as the characters i to p, are distributed as shown in Fig. 4, the correspondence table for these characters is shown in Table 1.

Table 1

Character i	Character i, Character j, Character n
Character j	Character j, Character n, Character ?
Character k	Character k
Character l	Character l
Character m	Character m
Character n	Character n

Character o	Character o, Character m
Character p	Character p

This correspondence table is really written by a character code of each character. The character ? means a special code to be assigned to a character not recognized.

If a character string to be retrieved is "character j + character m + character o" of three characters, the character j is replaced with a sum-set of characters j, n and ? and the character o is replaced with a sum-set of characters o and m.

Therefore, the retrieval is performed by using a sum-set of six character strings:

"characters j, m and o"

"characters j, m and m"

"characters n, m and o"

"characters n, m and m"

"characters ?, m and o"

"characters ?, m and m"

In the above manner, the information processing unit 2 broadens the range of the character string to be retrieved. The information processing unit 2 reads the character code information in the designated range from the storage device 3d to retrieve the character string matching any one of the character strings in the broadened range. The information processing unit 2 transfers the retrieved and extracted document information to the output unit 6. If

the output unit 6 is a CRT, the document information is displayed thereon, or if the output unit 6 is a printer, the document information is printed out.

As above, according to this embodiment, in retrieving a character string from character code information obtained through character recognition of document information input as image information, the target character string is replaced with a sum-set of the target character string and all character strings having a possibility of being erroneously recognized as the target character string, by referring to the correspondence table indicating a tendency of erroneous recognition inherent in a character recognition algorithm. In this manner, since the retrieve is performed by substantially broadening the retrieval range, even if the character recognition unit 5 erroneously recognizes a character, the occurrence frequency of retrieval omissions can be reduced considerably.

In the above embodiment, the retrieval target character string is replaced with a sum-set of the target character string and all other character strings likely to be erroneously recognized, by referring to the correspondence table. According to the main aspect of the invention, if character codes obtained by character recognition are used by retrieval only, erroneous recognition does not pose a serious problem on the assumption that recognition incompleteness is compensated. From this point of view, the range of the character string to be retrieved is substantially broadened to compensate for the recognition

incompleteness. Means for substantially broadening the range of a character string to be retrieved is therefore not limited to a particular means.

Effects of the Invention

As described so far, according to the invention, in retrieving a character string from character codes obtained through character recognition of a document, the range of the character string is substantially broadened so as to compensate for the character recognition incompleteness, so that an occurrence frequency of retrieval omissions can be reduced considerably.

4. Brief Description of the Drawings

Fig. 1 is a block diagram of an information storage apparatus according to an embodiment of the invention, Fig. 2 is a flow chart illustrating the operation of the information storage apparatus of the embodiment, Fig. 3 is a conceptual diagram illustrating ideal character recognition, Fig. 4 is a conceptual diagram illustrating incomplete character recognition, and Fig. 5 is a flow chart illustrating the operation of a conventional information storage apparatus.

2... information processing unit, 3... storage device, 4... code input unit, 5... character recognition unit.

Name of Agent: Attorney Akira KOKAJI and two others

FIG. 1

1... IMAGE INPUT UNIT, 2... INFORMATION PROCESSING UNIT, 3... STORAGE DEVICE, 4... CODE INPUT UNIT, 5... CHARACTER RECOGNITION UNIT, 6... OUTPUT UNIT

FIG. 2

(1)...START, (2)... READ DOCUMENT WITH IMAGE SCANNER, (3)... STORE DOCUMENT IN STORAGE DEVICE AS IMAGE INFORMATION, (4)... ENTER RETRIEVAL KEYWORD FROM KEYBOARD, (5)... STORE KEYWORD IN STORAGE DEVICE, (6)... RECOGNIZE CHARACTERS FROM IMAGE INFORMATION AND STORE CHARACTER CODES IN STORAGE DEVICE AT PREDETERMINED LOCATION AS AUXILIARY INFORMATION, (7)... REGISTRATION COMPLETED ?, (8)... ENTER RETRIEVAL RANGE FROM KEYBOARD, (9)... ENTER RETRIEVAL CONDITION FROM KEYBOARD, (10)... BROADEN CHARACTER STRING IN CONDITION EQUATION BY USING CORRESPONDENCE TABLE, (11)... RETRIEVE BY BROADENED RETRIEVAL CONDITION AND AUXILIARY INFORMATION IN RETRIEVAL RANGE, (12)... NARROW DOWN RETRIEVE RANGE ?, (13)... SET RETRIEVAL RESULT TO RETRIEVAL RANGE, (14)... OUTPUT RETRIEVE RESULT, (15)... END

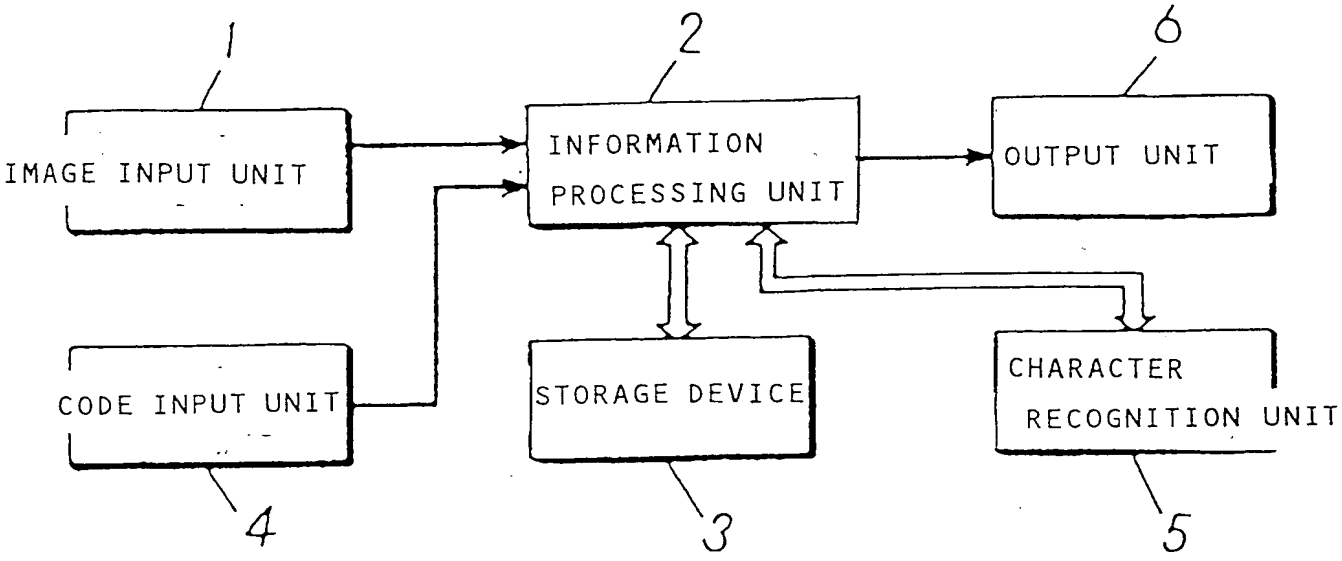
FIG. 5

(1)...START, (2)... READ DOCUMENT WITH IMAGE SCANNER, (3)... STORE DOCUMENT IN STORAGE DEVICE AS IMAGE INFORMATION, (4)... ENTER RETRIEVAL AUXILIARY INFORMATION FROM KEYBOARD, (5)... STORE AUXILIARY INFORMATION IN STORAGE DEVICE AT PREDETERMINED LOCATION, (6)... REGISTRATION COMPLETED ?, (7)... ENTER RETRIEVAL RANGE FROM KEYBOARD, (8)... ENTER

```
    RETRIEVAL CONDITION FROM KEYBOARD, (9)... RETRIEVE BY
    AUXILIARY INFORMATION IN RETRIEVAL RANGE, (10)... NARROW
    DOWN RETRIEVE RANGE ?, (11)... SET RETRIEVAL RESULT TO
    RETRIEVAL RANGE, (12)... OUTPUT RETRIEVE RESULT, (13)...
5   END
```

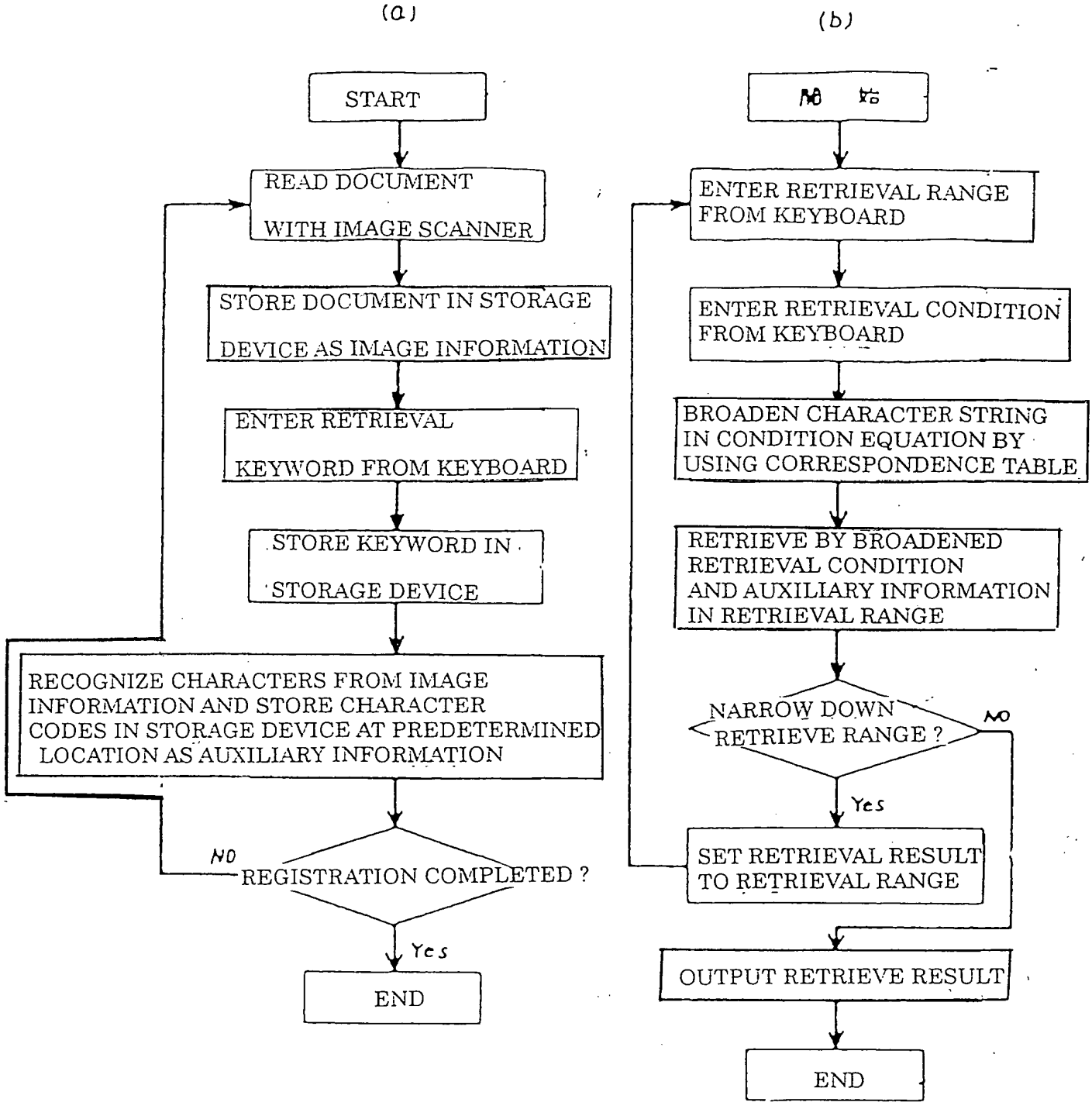
第 1 図

FIG. 1


1
2
6
IMAGE INPUT UNIT
INFORMATION PROCESSING UNIT
OUTPUT UNIT
CODE INPUT UNIT
STORAGE DEVICE
CHARACTER RECOGNITION UNIT
4
3
5

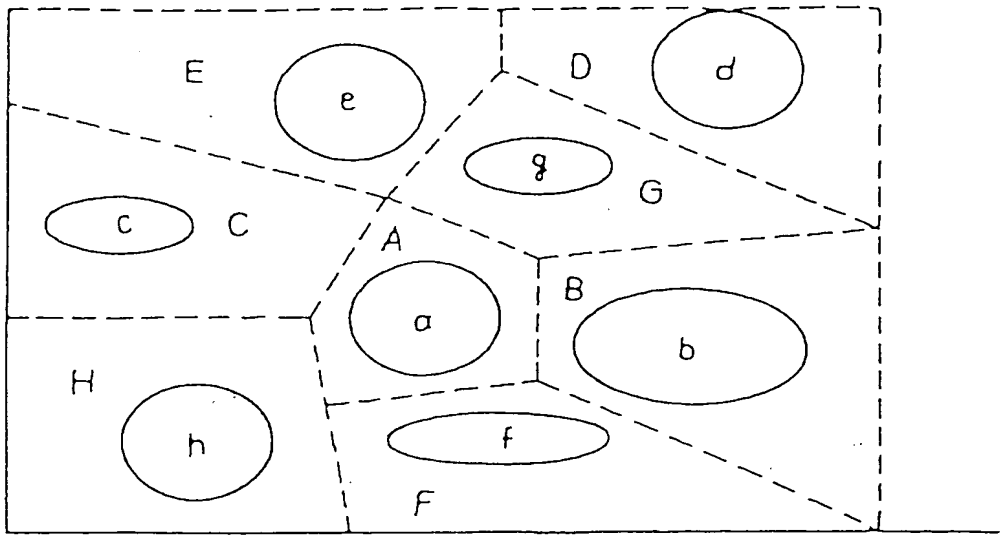
第 2 図

FIG. 2


(a)
START
READ DOCUMENT WITH IMAGE SCANNER
STORE DOCUMENT IN STORAGE DEVICE AS IMAGE INFORMATION
ENTER RETRIEVAL KEYWORD FROM KEYBOARD
STORE KEYWORD IN STORAGE DEVICE
RECOGNIZE CHARACTERS FROM IMAGE INFORMATION AND STORE CHARACTER CODES IN STORAGE DEVICE AT PREDETERMINED LOCATION AS AUXILIARY INFORMATION
NO
REGISTRATION COMPLETED ?
Yes
END
(b)
ENTER RETRIEVAL RANGE FROM KEYBOARD
ENTER RETRIEVAL CONDITION FROM KEYBOARD
BROADEN CHARACTER STRING IN CONDITION EQUATION BY USING CORRESPONDENCE TABLE
RETRIEVE BY BROADENED RETRIEVAL CONDITION AND AUXILIARY INFORMATION IN RETRIEVAL RANGE
NARROW DOWN RETRIEVE RANGE ?
NO
Yes
SET RETRIEVAL RESULT TO RETRIEVAL RANGE
OUTPUT RETRIEVE RESULT
END

【図 3】

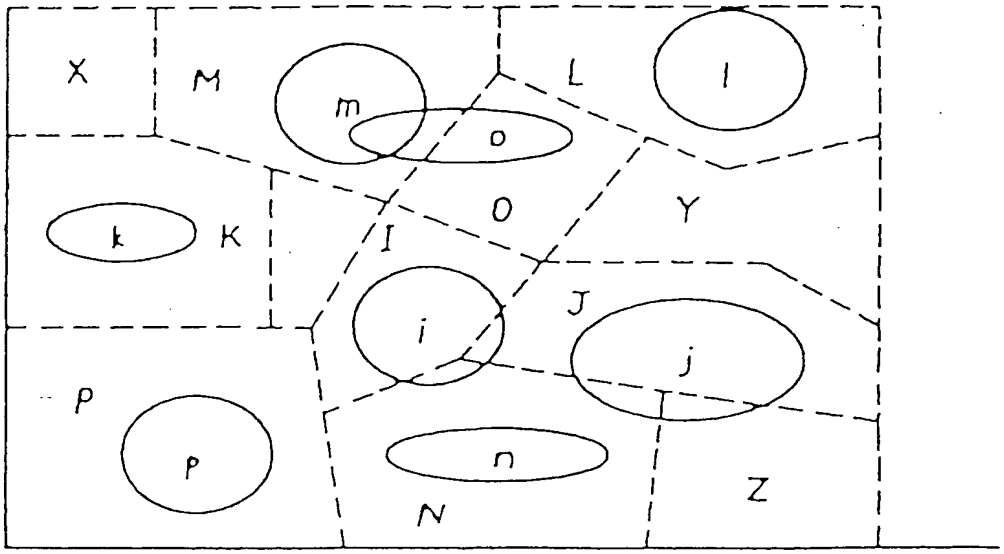
FIG. 3



CHARACTER PATTERN SPACE

【図 4】

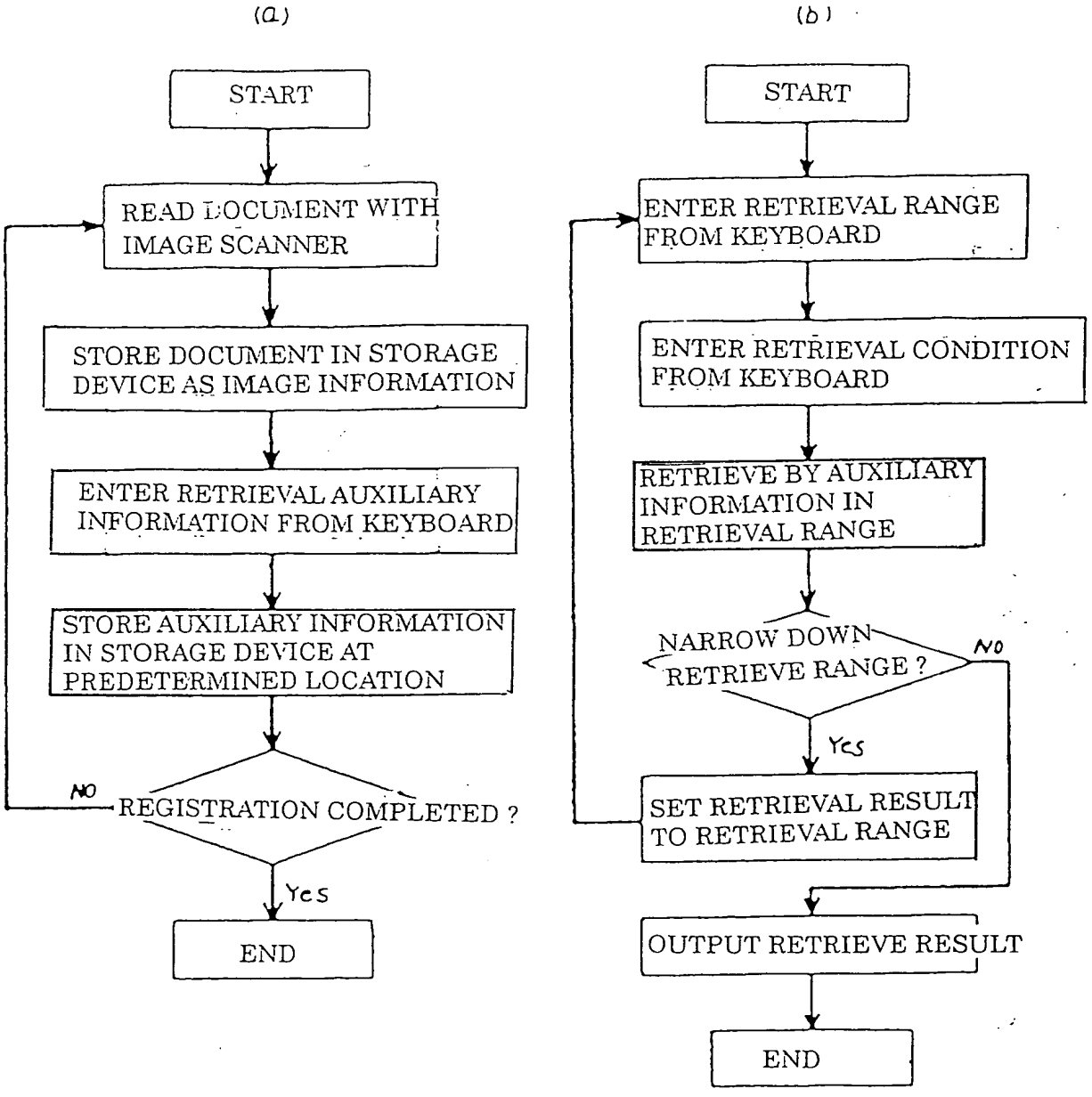
FIG. 4



CHARACTER PATTERN SPACE

第 5 図

FIG. 5


(a)
START
READ DOCUMENT WITH IMAGE SCANNER
STORE DOCUMENT IN STORAGE DEVICE AS IMAGE INFORMATION
ENTER RETRIEVAL AUXILIARY INFORMATION FROM KEYBOARD
STORE AUXILIARY INFORMATION IN STORAGE DEVICE AT PREDETERMINED LOCATION
REGISTRATION COMPLETED ?
NO
Yes
END
(b)
START
ENTER RETRIEVAL RANGE FROM KEYBOARD
ENTER RETRIEVAL CONDITION FROM KEYBOARD
RETRIEVE BY AUXILIARY INFORMATION IN RETRIEVAL RANGE
NARROW DOWN RETRIEVE RANGE ?
NO
Yes
SET RETRIEVAL RESULT TO RETRIEVAL RANGE
OUTPUT RETRIEVE RESULT
END